Finite sample weighting of recursive forecast errors

Article

Accepted Version

Finite Sample Weighting of Recursive Forecast Errors

Chris Brooks*

ICMA Centre, Henley Business School, University of Reading

Simon Burke

Department of Economics, University of Reading

Silvia Stanescu

Kent Business School, University of Kent

April 2015

THIS IS THE AUTHORS' ACCEPTED MANUSCRIPT OF AN ARTICLE FORTHCOMING IN THE INTERNATIONAL JOURNAL OF FORECASTING. THE DEFINITIVE VERSION IS AVAILABLE AT
http://www.journals.elsevier.com/international-journal-of-forecasting/

Abstract
This paper proposes and tests a new framework for weighting recursive out-of-sample prediction errors in accordance with their corresponding in-sample estimation uncertainty. In essence, we show how as much information from the sample as possible can be used in the evaluation of prediction accuracy by commencing the forecasts at the earliest opportunity and weighting the prediction errors. We demonstrate through a Monte Carlo study that when only a small sample is available the proposed framework can select the correct model from a set of candidate models considerably more often than the existing standard approach. We also show that the proposed weighting approaches result in tests of equal predictive accuracy which have much better size than the standard approach. An application to a set of exchange rate data highlights relevant differences in the results of tests of predictive accuracy based on the standard approach versus the framework proposed in this paper.

Keywords: forecast evaluation; forecast comparison; recursive model estimation; mean squared error; forecast weighting scheme.
JEL Classifications: C52, C53.
Acknowledgements: We thank two anonymous referees and the Handling Editor, Graham Elliott, for very useful comments that resulted in substantial changes to this paper. We are also particularly grateful to Kenneth Rogoff and Hali Edison for supplying the data used in the Meese-Rogoff study. We also acknowledge the useful comments of Mike Clements and seminar participants at CORE, Université Catholique de Louvain; the Norwich Business School; and the College of Business, Economics and Law, Swansea University.
* Contact: Corresponding author: Chris Brooks, ICMA Centre, University of Reading, Whiteknights, Reading RG6 6BA, UK; tel +44 118 378 7809; fax: +44 118 931 4741; e-mail: C.Brooks@reading.ac.uk

1. Introduction

The issue of forecast evaluation is key in many assessments of model adequacy, and consequently has received considerable attention. As Diebold (2013) notes, forecast evaluation is in practice rarely the object of interest in its own right but rather, it is more often conducted as a route to evaluating the comparative accuracy of competing models. In such circumstances, one of two approaches may be adopted – either to focus entirely on in-sample model estimation over all observations available, or to split the data into an in-sample estimation part and a separate out-of-sample forecast portion with the evaluation then taking place entirely based on the latter.

In the out-of-sample forecasting literature, three alternative frameworks for conducting performance tests have been employed: constant coefficients, rolling windows of constant size, and recursive forecasting. The choice between these approaches is sometimes motivated by the particular forecasting application being considered but more often is entirely arbitrary. Examples of studies using each approach include Ashley, Granger and Schmalensee (1980) (constant coefficients), Cheung, Chinn and Pascual (2003) (rolling) and Faust, Rogers and Wright (2004) (recursive). West (2006) argues that the constant coefficient approach is preferred in cases where the (re-)estimation process is impossible to take into account while the rolling window alternative may be preferable when there are structural breaks or regime shifts in the series. One could argue intuitively that recursive forecasting would be preferable in many situations where the sample size is small since it makes use of all of the information available to the forecaster at that point in time. Faust, Rogers and Wright (2004) compare the mean squared errors (MSEs) obtained using a constant coefficients forecasting scheme with those resulting from a recursive framework, and find that the recursive one almost always produces lower MSEs and that the difference between the two forecasting methodologies is statistically significant in some of their samples.

In a point prediction application, it is often the case that the forecaster aims to find the model from among a set of candidates that minimises the value of a previously determined loss function, conditional on the information available at the time that the forecast is made. Such prediction comparison studies have been termed "forecasting horse races". The key innovations in this literature over the past two decades have almost invariably concerned either the models employed, which have become more sophisticated,[1] or the loss functions adopted, which have increasingly tended towards economically relevant measures. Yet the forecast evaluation framework itself has scarcely warranted a mention and the vast majority of studies still base this on an ad hoc rule of

[1] As Diebold (2013) notes, perhaps unsurprisingly it is usually the new horse in the stable that wins the race.

thumb whereby a fixed proportion of the data is used for in-sample model estimation and the remainder retained for out-of-sample evaluation where the predictions are compared with actual values.

Given a fixed overall quantity of data, it would be possible to have a short in-sample period with a long evaluation period, two samples of roughly equal length, or a long in-sample period and only a small hold-out sample. Clearly there is a trade-off to be made here. If the in-sample estimation period is too short, parameter uncertainty will be high, leading to forecast imprecision; on the other hand, if the out-of-sample estimation period is too short, even if the models are estimated with reasonable accuracy, there will be too few forecasts with which to compare the actual values and hence the metrics for forecast evaluation (e.g., the out-of-sample MSE) will be noisy and unreliable.[2] There are examples of studies using low, medium and high proportions of the data out-of-sample and in the vast majority of research it is not at all clear how this choice has been made – it appears to be subjective, capricious and tilted towards a long estimation period with a consequent short evaluation part.[3] West (1996) shows that parameter estimation will asymptotically not affect the outcome of tests of the equivalence of means squared errors from non-nested models with conditionally homoscedastic errors and this intuitively leads to a preference for using a relatively large in-sample estimation period. When the in-sample period is relatively large (say, 90% of the available data or more), the impact of parameter estimation error can probably be ignored (West, 2006); however, in the context of nested models or when the number of observations is small, it is debatable whether the asymptotic irrelevance will still hold.[4]

An interesting puzzle in the empirical economic forecasting literature is that there exist numerous studies showing that a particular variable or set of variables possesses in-sample predictability which cannot be retained in out-of-sample tests. The conventional explanation has been that the in-sample forecasting ability was illusory and a consequence of data mining. An alternative explanation, which Inoue and Kilian (2005) support through asymptotic theory, is simply that out-of-sample tests in many instances lack sufficient power to detect predictability that is actually

[2] Ashley (2003) shows that to demonstrate one forecasting model to be statistically significantly better than another would typically require more than 100 out-of-sample observations. Yet when quarterly or even monthly data are employed with in-sample estimation windows of conventional length, fewer out-of-sample data points than this are commonly available.

[3] For example, West's (2006, p.106) review paper explicitly takes the split point "as given" with no further discussion.

[4] A recent study that directly addresses the issue of the position of the split into in-sample and out-of-sample periods is by Hansen and Timmerman (2012). Building on earlier work by McCracken (2007) and Clark and McCracken (2001, 2005a), Hansen and Timmerman develop an approach that modifies the p-values of tests of the null hypothesis of no predictability so that they become robust to sample-split-induced data-mining. A conceptually similar modification, albeit different in detail, is proposed by Rossi and Inoue (2012) and further comparisons of the power of tests for the differences between out-of-sample forecasts can be found in Busetti and Marcucci (2013).

present in the data – an inevitable consequence of the loss of data because of the sample splitting. As a result, they advocate the sole use of in-sample *t*- or *F*-tests to compare between nested models (with no out-of-sample analysis).[5]

When the model-building and forecasting exercise occurs in the context of small samples of data having non–standard features or in the presence of model misspecification, however, the asymptotical results concerning the desirability of in-sample testing may no longer apply. Inoue and Kilian (2006) note two circumstances where out-of-sample testing may be favoured relative to in-sample model selection with SIC: when comparing between non-nested models in the context of autocorrelated data, and when comparing nested models when the true model is the larger one. In this case, the smaller (incorrect) model will suffer less from parameter estimation error and this reduction in variance will more than compensate for the additional forecast error bias arising from the use of the wrong model. So in-sample analysis may be preferable in general but this conclusion could vary considerably dependent on the context and precise nature of the data. A further reason to prefer out-of-sample testing to a pure in-sample evaluation is that a researcher might be interested in how a model performs in prediction at a particular point in time rather than on average.[6]

While it is perhaps difficult to generalise, it is quite frequently the case that around two thirds of the sample or more are used for initial in-sample model estimation, leaving the remaining third or less as a hold-out sample.[7] However, in cases where the overall amount of data available is small, this is potentially very wasteful. In this study, we propose an approach which is conducted within what we term a fully recursive framework where out-of-sample forecasting begins at the earliest possible opportunity. The initial forecasts from such an approach are likely to be imprecise, having high error variance, because their construction will be based on model parameters estimated using very small samples. On the other hand, the conventional approach using two thirds or more of the data for in-sample estimation would have implicitly assigned such forecasts a weight of zero. We argue that this is not ideal since the early forecasts, while likely very noisy, will still contain useful

[5] Relatedly, Clark and McCracken (2005b) demonstrate that an in-sample F-test of predictive ability is likely to be more powerful than out-of-sample tests. Schwarz's information criterion (SIC) could instead be used to penalise additional parameters within the model evaluated in-sample, and will deliver the correct model asymptotically with probability one when it is in the choice set. Moreover, SIC will still consistently select the best approximating model even when all in the choice set are misspecified (see Inoue and Kilian, 2006). While the penalty term in SIC may help to weed out spurious predictability arising from data-mining in such contexts, as Diebold (2013) notes, it will not help if the researcher wishes to compare between two (non-nested) models containing the same number of parameters, one of which has arisen as a result of data-mining.

[6] Giacomini and Rossi (2010) develop a method for testing precisely this 'local' forecasting power.

[7] Indeed, in their comparison between in-sample versus out-of-sample model selection approaches, Inoue and Kilian (2006) use a 90%-10% split in their base case Monte Carlo simulation, although they note that such a choice is likely to maximise the differences between the performances of the two approaches.

information regarding the accuracy of the model in prediction. Instead, we propose giving these early predictions lower weight in the forecast error aggregation function with the weights increasing as the in-sample recursive estimation period grows and therefore the ex ante forecast error variance declines. In essence, bringing the start of the hold-out sample evaluation period forward as early as possible removes user-discretion regarding the sample split position and therefore the opportunity to data mine (at least along that dimension). We show how tests of model adequacy can be conducted in a standardised framework that varies from one empirical application to another as little as possible and we investigate the efficacy of the approach using a Monte Carlo study. We show that, when only a small sample of data is available, severe distortions in the ranking of competing forecasting models may appear when the out-of-sample part of the data begins late in a recursive scheme.[8]

2. Developing a Generalised MSE[9]

We start with a linear model, which is being estimated using recursive OLS based on a sample of T observations, as follows:

$$y_\tau = x_\tau' \beta + \varepsilon_\tau, \ \tau = 1, 2, \ldots, T, \tag{1}$$

where x_τ and β are $k \times 1$ vectors of explanatory (exogenous) variables and parameters respectively, y_τ the scalar process to be forecast and ε_τ a scalar disturbance sequence. Competing forecasting models will therefore be of the form:

$$y_\tau = x_\tau^{(i)\prime} \beta^{(i)} + \varepsilon_\tau^{(i)}, \ \tau = 1, 2, \ldots, T, \tag{2}$$

where $x_\tau^{(i)}$ and $\beta^{(i)}$ are $k^{(i)} \times 1$ vectors corresponding to the i^{th} of m competing models (i.e. $i = 1, \ldots, m$). We are only concerned here with 1-step ahead forecasts, which, in this framework are of the form:[10]

$$y_{t+1|t}^{(i)} = x_{t+1}^{(i)\prime} b_t^{(i)} \tag{3}$$

where $b_t^{(i)}$ is the OLS estimate of $\beta^{(i)}$ in (2) based on the sample $\tau = 1, 2, \ldots, t$; it is a $k^{(i)} \times 1$ vector given by:

[8] We should state at the outset that we do not provide a formal mathematical proof of the optimality of the forecast weighting and in that sense it is a somewhat ad hoc modification. However, we argue that the approach has strong intuitive appeal and, as we show below, works well for small samples.

[9] The methodology and weighting schemes proposed in this section are applicable for alternative forecast error aggregation measures as well, such as, for example, the mean absolute error (MAE). However, we do not consider these further due to space constraints.

[10] Note that, following numerous existing empirical studies (including, most notably, Meese and Rogoff, 1983) we assume throughout the paper and in the Monte Carlo study that the values of x_{t+1} are known. Of course, in any real application it is likely to be the case that if they are exogenous variables these must be forecast as well.

$$b_t^{(i)} = \left(X_t^{(i)\prime} X_t^{(i)}\right)^{-1} X_t^{(i)\prime} Y_t \tag{4}$$

with

$$Y_t = \left(y_1 \quad \cdots \quad y_t\right)', \ t \times 1; \tag{5}$$

$$X_t^{(i)} = \left(x_1^{(i)} \quad \cdots \quad x_t^{(i)}\right)', \ t \times k^{(i)} \tag{6}$$

Define the OLS one-step ahead forecast error based on the recursive sample $\tau = 1, 2, \ldots, t$, again for each model i, as:

$$v_{t+1|t}^{(i)} = y_{t+1} - x_{t+1}^{(i)\prime} b_t^{(i)}, \quad \text{scalar}; \tag{7}$$

The common practice in the literature is that forecast errors resulting from a recursive scheme are given equal weight in the model evaluation criterion/loss measure,[11] with the squared error loss or quadratic loss criterion being the criterion most often used in practice. When the mean forecast error is zero (i.e. unbiased forecasts), the expected squared forecast error is equal to the variance of the forecast error. Hence, by minimising the expected squared forecast error, the forecaster choosing among competing unbiased forecasts selects the minimum variance forecast. If the objective of the forecaster is the minimisation of the expected value of the squared forecast error (i.e. a MSE criterion), then in practical applications, one minimises the sample counterpart of this expectation, the sample mean. This is how the equal weighting of observations – weight equal to one over the total number of forecast errors – appears in the MSE criterion. The empirical (sample) mean squared error is obtained by equally weighting all squared forecast errors (weight $1/(T\text{-}k)$), where $k = k_{\max} = \max_i \left(k^{(i)}\right)$ and hence $T\text{-}k$ is the number of (out-of-sample) forecasts in our framework). For model i we write:

$$MSE = \sum_{t=k}^{T-1} \frac{1}{T-k} \left[v_{t+1|t}^{(i)}\right]^2 \tag{8}$$

We can now generalise the well-known expression in (8) as:

$$GMSE = \sum_{t=k}^{T-1} w_t^{(i)} \left[v_{t+1|t}^{(i)}\right]^2 \tag{9}$$

where $0 \le w_t^{(i)} \le 1$ and $\sum_{t=k}^{T-1} w_t^{(i)} = 1$.[12] It is easily noticeable that the usual MSE is just a special case in the specification in (9), with $w_t = w = 1/(T\text{-}k)$ for all t. Intuitively, a sound criterion for forecast

[11] Strictly, loss functions are usually defined as a function of expectations over a random variable; however, here we estimate them using their sample counterparts based on actual data. So in order to minimise confusion, we refer to these objects as "model evaluation criteria" in this paper.

[12] Note that in equation (8), the divisor remains inside the summation sign purely for ease of comparison with (9) although it does not vary with t.

evaluation might take into account not only the (ex post) accuracy of the forecasts – i.e. how large the (squared) forecast errors were ex post – but also the (ex ante) uncertainty of different forecasts.

We note that while the uncertainty is different for different forecasts when employing recursive forecasting, the standard, equally weighted *MSE* criterion weighs all forecast errors equally. We propose a number of alternative weighting schemes which take into account this varying uncertainty. Unless otherwise specified, for all weighting schemes below we use:

$$w_t^{(i)} = \frac{\tilde{w}_t^{(i)}}{\sum_{t=k}^{T-1} \tilde{w}_t^{(i)}} \tag{10}$$

so that the weights always sum to unity with the various schemes in Sections 2.1-2.5 specifying different $\tilde{w}_t^{(i)}$. It is necessary to rescale the weights in this manner in order to ensure that a model with a persistently large forecast error variance and therefore persistently small $\tilde{w}_t^{(i)}$ does not have a lower *GMSE* purely as a result of this.

2.1 A weighting scheme based on the variance of forecast errors

A first weighting scheme we propose is:

$$\tilde{w}_t^{(i)} = \frac{1}{Var\left(v_{t+1|t}^{(i)}\right)} \tag{11}$$

If model *i* is correctly specified then it can be shown (see Philips and Harvey, 1974; Brown, Durbin and Evans, 1975; and Harvey, 1988) that the one-step ahead forecast errors $v_{t+1|t}^{(i)}$ have zero mean and variance equal to:

$$Var\left(v_{t+1|t}^{(i)}\right) = \left(\sigma^{(i)}\right)^2 f_{t+1}^{(i)} \tag{12}$$

and are uncorrelated, where $\left(\sigma^{(i)}\right)^2$ is the variance of the disturbance process corresponding to model *i,* and:

$$f_t^{(i)} = 1 + x_t^{(i)\prime}\left(X_{t-1}^{(i)\prime} X_{t-1}^{(i)}\right)^{-1} x_t^{(i)} \tag{13}$$

It can further be shown (see Harvey, 1990, p. 55) that if model *i* is correctly specified and the forecasts for model *i* begin at observation $k^{(i)}+1$, then:

$$\sum_{t=k^{(i)}}^{T-1} \frac{\left(v_{t+1|t}^{(i)}\right)^2}{f_{t+1}^{(i)}} = RSS_T^{(i)} \tag{14}$$

where $RSS_T^{(i)}$ is the residual sum of squares from in-sample estimation of the model using all observations T. Therefore, assuming correct specification and using the properties of recursive residuals, $v_{t+1|t} \Big/ \sqrt{f_{t+1}}$, for $k^{(i)} = k$ we get that:

$$GMSE = \frac{RSS_T^{(i)}}{\sum_{t=k}^{T-1} \frac{1}{f_{t+1}^{(i)}}} \tag{15}$$

The proposed criterion can thus be expressed as a combination of in-sample fit (corresponding to the estimation performed for the entire sample T) and variance of the forecast errors: the higher are these variances (i.e. the f_t), the higher the *GMSE*. Furthermore, if $k^{(i)}<k$, then the expression in (15) becomes:

$$GMSE = \frac{RSS_T^{(i)} - RSS_{k-1}^{(i)}}{\sum_{t=k}^{T-1} \frac{1}{f_{t+1}^{(i)}}} \tag{16}$$

If the data generating process is a random walk or a random walk with drift, $Var\left(v_{t+1|t}^{(i)}\right) = \left(\sigma^{RW}\right)^2$, a constant and hence *GMSE* = *MSE* in this case, although the MSE will differ across the two models since the forecasts themselves will differ.

The purpose of this paper is to study the implications (with regards to ranking of competing models) of weighing forecast errors resulting from a recursive scheme in accordance with their (varying) inaccuracy of estimation rather than assigning them equal weights as is usually done in the forecasting literature. Thus, at each point in time, the weight would be a function of the model uncertainty at that point, which would be declining as the number of observations used in the in-sample estimation increased. The weighting scheme proposed above weights the squared one-step ahead squared forecast errors by their (standardised) variance, $\frac{1}{f_{t+1}^{(i)} \sum_{t=k}^{T-1} \frac{1}{f_{t+1}^{(i)}}}$.

In other words, instead of defining the model evaluation criterion based on forecast errors, we define it on (partially) standardised forecast errors. We are then still optimising based on the expected quadratic model evaluation criterion and its sample counterpart, but the variable (on which the model evaluation criterion depends) changes – from error to standardised error. This is equivalent to changing from equal weighting of the squared forecast errors to a different scheme with weights based on the inverse of the variance of the forecast error (or a proxy for it). Hence,

since the forecast errors have (time-)varying variance in a recursive scheme, we argue that it is not the *MSE* criterion that we ought to employ, but rather the *GMSE* (in other words, the mean squared *standardised* error). In the following three sections, we now describe alternative weighting functions *f* that are also somewhat ad hoc in nature but considerably more intuitive and easier to implement.

2.2 An Alternative Weighting Scheme Based on the Number of In-sample Observations

For an AR(1), Clements (2005) gives an approximate expression for the variance of the (*h*-step ahead) forecast error which shows that the effect of uncertainty coming from parameter estimation on the (approximate) variance of forecast errors is proportional to 1/(sample size).[13] Thus parameter uncertainty is large when the *estimation sample size* is small and vice versa. Consequently, we expect forecast error variances to decrease as the in-sample period increases when a recursive scheme is used, thus increasing the precision of the forecasts. This suggests another weighting scheme that could be employed – namely, weighting the forecast error by the square root of the inverse of the sample size used for estimation (so that the squared forecast errors are weighted by the inverse of the sample size); the expression for $\tilde{w}_t$ now becomes:[14]

$$\tilde{w}_t = \frac{1}{1+\frac{1}{t}}, \; t = k, \ldots, T\text{-}1 \qquad (17)$$

2.3 An Alternative Scheme Based on Log-time weighting of the Forecast Errors

This is another heuristic approach based on the same arguments as used in the previous sub-section. In (17), the squared forecast errors are weighted by the number of in-sample observations, and thus the weighting function is effectively linear in time. Such a function may not have the desired shape, particularly when *t* is large, since it will give very small weight to the initial squared forecast errors. An alternative weighting function that we employ for comparison replaces *t* in the denominator in (17) by (1+log(*t*)). While this function is still monotonically increasing in *t*, it increases at a decreasing rate.

2.4 An alternative Weighting Scheme Based on the Standard Error of the Regression

[13] See Clements (2005), p.11.

[14] A slight variation on this would be to use *t-k* in the divisor rather than *t* so that a degrees of freedom adjustment is made. However, we do not pursue this. Also, we drop the superscript (*i*) in equation (17) since these set of weights do not depend on the model specification.

A final alternative that we also examine replaces the f_t by the standard error of the regression (SE(regression) hereafter) estimated on the sample up to observation t, ($\tau = 1,2,\ldots,t$). The weights would now be (for the forecast made at time t, for time t+1) given model i:

$$\tilde{w}_t^{(i)} = \frac{t - k^{(i)}}{RSS_t^{(i)}} \tag{18}$$

3. The Monte Carlo Simulation and its Results

3.1 Design of the Monte Carlo

In testing how well the various approaches perform, we construct a Monte Carlo framework with six competing models. Separately, we make each of the models the correct one and thus it is the data generating process in that case. Then each of the six models is estimated and a set of one-step ahead forecasts is produced with the forecasts being evaluated on the basis of their MSE and GMSEs described above. The Monte Carlo is run with 10000 replications, and we measure the proportion of times that each model is selected as being the best (i.e. the one with the lowest MSE/GMSE) for each DGP. The approaches that we use are:

1. Standard recursive estimation from $k_{\max}$+1 and forecasting with no weighting.
2. Fixed 2/3 in-sample and 1/3 out-of-sample split with a recursive window – we might term this the standard framework for forecast evaluation.
3. Recursive estimation from $k_{\max}$+1 with various weighting schemes. The first of these is the 1/*f*-weighting but we also employ, as discussed above, the more ad hoc weighting schemes including t, 1+log(t), 1/SE(regression).

Thus, overall, we have six weighting schemes. We initially define the best approach from among these six as the one that selects the correct model as preferred the highest proportion of times. We employ five sample sizes: 15, 25, 50, 100 and 250 in total, and we also allow for a burn-in for each series generated under the DGP of 100 observations. The error distribution is a standard normal and thus in the equations below, u_t ~NIID(0,1). The models we employ are intended to cover a range of empirically relevant specifications from the economics and finance literature. As far as possible we try to ensure that the models are not nested within one another and that they cover different types of regression models that might be used in financial and economic forecasting. We therefore include the random walk (with drift) which is the competitor to beat in many forecast model races, as well as a time series (autoregressive) model, representing another class a models often successfully employed in financial forecasting. Three different specifications of structural models are also

considered, alongside what we term the 'piecewise constant mean' model which allows for a structural break.

Random walk with drift

$$y_t = 0.5 + y_{t-1} + u_t \quad (19)$$

AR(4)

$$y_t = 0.5 + 0.9y_{t-1} - 0.1y_{t-2} + 0.2y_{t-3} - 0.6y_{t-4} + u_t \quad (20)$$

Linear structural model

$$y_t = 0.5 + 0.8x_t + u_t \quad (21)$$

(We assume the true value of x in the forecast with $x_t \sim$NIID(0,1)).

Quadratic structural model

$$y_t = 0.5 + 0.8x_t^2 + u_t \quad (22)$$

Multiplicative Structural model

$$y_t = 0.5 + 0.8x_t z_t + u_t \quad (23)$$

(We assume the true values of x and z in the forecast, x_t, $z_t \sim$NIID(0,1) and the correlation between x and z is a constant ρ, such that $z_t = \rho x_t + \sqrt{1-\rho^2}v_t$, where u_t, v_t and x_t are separate NIID(0,1), $\rho = 0.2$).

Piecewise constant mean model

$$y_t = 0.5 + 0.8D_t + u_t \quad (24)$$

(where $D_t = 0$ for the first $T/2$ observations and 1 thereafter).

So to give an illustration of how this works, suppose that T=25 and that the data generating process is a random walk with drift. We thus construct 126 data points (including 100 for burn in and one for the pre-sample value) of the model according to equation (19). We then estimate all of the models above (even though we know that five of these models are wrong) using the first five observations ($k_{\max}$), with the forecasts beginning at observation 6 for each of the weighting functions except the standard one based on an arbitrary 2/3-1/3 sample split. For the standard approach, the first 17 observations are used for in-sample estimation, with the forecasts effectively beginning at observation 18. We then produce the forecasts from all seven models and compute their MSEs in each case, weighting according to the various schemes described above. The model with the lowest MSE for a given weighting scheme is the chosen one for that replication. We then re-run the experiment a further 9999 times and note the model selected in each case. A good weighting criterion should select the random walk with drift, which is the correct model and constitutes the DGP in this case, a relatively high proportion of the time.

Figure 1 presents a graphical illustration of how the weighting schemes work for an overall sample size of T=30. Given that k-max in our setup is five, the one-step ahead forecasting begins at observation six and then continues in a recursive manner with an additional data point added until the end of the sample is reached.[15] In the figure, as in the Monte Carlo experiments and when the actual data are used, all of the weights for a given evaluation framework are normalised to sum to one. The standard approach based on using two thirds of the data for in-sample estimation will only begin forecasting from observation 21, giving a high but equal weight to the remaining forecasts and a zero weight to all forecasts before that. All of the other approaches give increasing weight to the forecasts as the estimation sample increases from left to right in the figure. For the deterministic schemes: t and 1+log(t), the increase is monotonic and linear for the t but increasing at a decreasing rate for 1+log(t). The remaining two schemes – the 1/f and 1/SE(regression) weightings – generally increase from left to right as the sample size increases and the amount of uncertainty falls; however, this increase is not monotonic as the addition of an outlying observation on y to the sample will increase rather than reduce parameter uncertainty. The other approach we employ, as discussed above, is what we term the fully recursive one, where the forecasting begins at observation six but with equal weighting (of 0.04 in this example) of forecast errors thereafter.

Intuitively, we would expect the standard 2/3-1/3 approach to perform poorly in general for the small sizes that are focus in this paper, due to the inefficient initial use of the 2/3 of the data for in-sample estimation and the equal weighting of forecast errors whatever the parameter uncertainty thereafter. Similarly, we would expect the fully recursive approach to perform poorly since the initial forecast errors early in the sample are given equal weighting in the model evaluation criterion but are likely to be highly inaccurate due to the very small sample size used in model estimation.

3.2 Statistical Comparisons of Forecast Accuracy

Formal tests for equality of predictive accuracy are commonly employed in the literature since it is often considered desirable to know whether the differences between competing models are statistically significant, and this information cannot be gleaned from a pure 'first past the post'

[15] In fact, a slight caveat is required here. When $t=k_{max}$, there are no degrees of freedom and so while a forecast can be produced for t+1, the standard error of the regression will be zero (since the model will fit perfectly). Hence the weighting function based on this will set a zero weight to the first forecast (i.e. for observation k_{max}+1).

evaluation.[16] By far the most popular such test is due to Diebold and Mariano (1995) – *DM* hereafter – who provide an asymptotic framework for comparing a sequence of *h*-step ahead forecasts from two models.[17] The null hypothesis is of equal predictive accuracy across the two models. Their approach is as follows:

(i) Define a metric, $L(.)$, to be applied to each forecast (in our case each one-step ahead forecast) for each of two models.

(ii) Take the difference between the two.

(iii) Average these differences over all the recursive forecasts produced.

(iv) Test whether this average difference is significantly different from zero estimating the variance of the difference using a long-run variance estimator.

As above, the notation (i) as a superscript is used to denote a quantity relating to the i^{th} model. The one-step ahead forecast from model i is based on sample τ = 1, 2, …, t and $t = k', \ldots, T-1; k' = \max(k^{(1)}, k^{(2)})$. The metric for our generalised mean squared error is given by

$$L_{t+1}^{(i)}(\hat{y}_{t+1}^{(i)}) = \left[\tilde{v}_{t+1}^{(i)}\right]^2. \tag{25}$$

where $\left[\tilde{v}_{t+1}^{(i)}\right]^2 = w_t^{(i)}\left[v_{t+1|t}^{(i)}\right]^2$.

The difference between the metrics is

$$d_{t+1} = \left[\tilde{v}_{t+1}^{(1)}\right]^2 - \left[\tilde{v}_{t+1}^{(2)}\right]^2. \tag{26}$$

The average difference is

$$\bar{d} = \frac{1}{T-k'}\sum_{t=k'}^{T-1} d_{t+1}. \tag{27}$$

Then the *DM* statistic is

[16] However, Inoue and Kilian (2006) argue that tests of equality of forecast accuracy are unnecessary since they will not affect the decision about which model to use – whatever the outcome of the test in terms of whether the null is rejected or not, the most accurate model will always be used.

[17] An alternative framework would be to construct forecast encompassing tests along the lines of, for example, Clements and Hendry (1993) or Harvey, Leybourne and Newbold (1998). However, we do not pursue this line of enquiry here.

$$DM = \frac{\bar{d}}{\left[\Omega(\bar{d})/T\right]^{1/2}}, \tag{28}$$

where $\Omega(\bar{d})$ is the long-run variance estimator of $\bar{d}$ employing the autocovariances of the d_t sequence.[18]

Diebold and Mariano's test was derived in the context of no parameter uncertainty, which in practice is tantamount to a comparison of pure forecasts. Diebold (2013) argues that the majority of studies employing the *DM* test have done so in the context of comparisons of forecasting accuracy in pseudo-out of sample environments, which is not what the test was intended for. The standard approach does not allow for the impact of estimation error on the tests of equality of mean squared prediction errors. However, recent research has examined the role that parameter uncertainty may play in affecting the properties of the test, and has proposed modifications or a slightly different approach accordingly (see, for example, McCracken, 2000; Clark and McCracken, 2001; West and McCracken, 1998; Clark and West, 2006; Clark and West, 2007). Inoue and Kilian (2005) advocate the derivation of bootstrapped critical values, although Diebold (2013) argues that this is unnecessary and that the tabulated critical values will be sufficient.[19] A particular issue arises in the comparison of the forecasts from nested models since the forecasts will converge asymptotically and are likely to be cross-correlated even in finite samples, resulting in the test being severely undersized and lacking power if standard critical values are employed (see Clark and McCracken, 2001; Clark and West, 2006; Busetti and Marcucci, 2013).

Nonetheless, it seems clear that the *DM* test in its earliest form is still by far the most prevalent in empirical work and Diebold (2013) argues that the assumption behind the original test can "often be credibly evoked even when comparing estimated models". Harvey, Leybourne and Newbold (1997) show through a Monte Carlo study that when the number of predicted data points exceeds around 30, the standard Diebold-Mariano test has acceptable size when the forecast horizon, *h*, is only one period; however, its performance rapidly deteriorates as *h* increases. *DM* in its standard form is therefore the focus of our attention in comparisons of forecast accuracy. A particular advantage is that it does not require forecast unbiasedness and can be applied in contexts other than quadratic model evaluation criterions (Harvey, Leybourne and Newbold, 1997).

[18] Harvey, Leybourne and Newbold (1997) develop a finite sample corrected version of this statistic.
[19] It has also been argued that the modifications are "burdensome to compute" (see Hansen and Timmermann, 2013, p.1, who again effectively advocate a return to in-sample testing).

Since the recursive one-step ahead forecast errors are not autocorrelated, it seems possible that the autocovariances of d_{t+1} might be zero, in which case the long run variance reduces to the standard sample variance estimator of the mean. Optimal multi-step ahead forecasts are almost certain to be serially correlated due to the overlapping horizons but we side-step this serious issue by focusing only on one-step ahead forecasts.[20]

3.3 The Simulation Results

The core results of the study – those of the Monte Carlo experiments – are presented in Tables 1 to 5 for total sample sizes, T, of 15, 25, 50, 100, and 250, observations respectively.[21][22] The cell entries are the percentages of the 10000 replications that each model is selected given a specific forecasting and weighting scheme and each row will sum to 100% (with possible slight discrepancies due to rounding errors). The tables all have six panels, each one for a different data generating process. In these tables, the results in Panel A show the percentage of times that each model is chosen when the data generating process is the random walk with drift, the Panel Bs show the percentage of times each model is chosen when the DGP is an AR(4), and so on. Thus entries on the leading diagonal of each table (column 1 in Panel A, column 2 in Panel B, ...) represent correct model selections and any off column-diagonal entries represent incorrect selections.

If we begin the analysis with T=15 in Table 1, it is evident that this represents an extremely small sample, perhaps of the order that might be employed when only annual data are available. Nonetheless, aside from the standard approach, all of the forecasting and weighting frameworks perform reasonably well. In Panel A the "1/f" weighting scheme selects the correct model (the random walk with drift) almost 95% of the time. Most of the other recursive approaches show around this level of accuracy while the 2/3-1/3 approach with no forecast error weighting can only identify the correct model around half of the time. Broadly the same findings hold when the data generating process is an AR(4) or a structural model. All approaches have difficulty picking out the piecewise linear model, with only approximately 50% correct identification for the recursive

[20] Under the null of equal predictive accuracy, *DM* has an asymptotic standard normal distribution. However, for this asymptotic result to hold, Diebold and Mariano (1995) state a (sufficient but not necessary) assumption that is important for our considerations: they assume that the difference d_{t+1} is covariance stationary. We note that in a recursive scheme the (un-weighted) one-step ahead forecast errors have time-varying variance; it is therefore likely that the squared loss differential, d_t, will also have non-constant variance which would contradict the stationarity assumption. Using the first weighting scheme that we propose in Section 2.1 would, however, resolve this problem.

[21] As a referee points out, at the smallest of these sample sizes, it is possible that all of the methods will have some difficulty in identifying the best model since the estimation errors will be very large.

[22] We also ran the simulations for T = 500 and 1000 observations, but the asymptotic behaviour is clearly already becoming evident by T = 250 and therefore we do not present results for the larger samples to preserve space.

schemes and a 22% hit rate for the 2/3-1/3 approach. Overall, the recursive approaches always outperform, and quite palpably, the 2/3-1/3 split. Regarding the recursive approaches, with the exception of the final model (the piecewise constant), there is always at least one weighting scheme which outperforms the non-weighted recursive approach, albeit mildly.

Increasing the total sample size to 25 observations (still a very short time-series) in Table 2 already results in a substantial improvement in evaluation performance across all models, although the 2/3-1/3 approach is still inferior to the others. Most of the models are correctly identified around 80-90% of the time by all criteria except for 2/3-1/3, which typically manages less than 70% accuracy. This standard approach appears to have particular difficulty identifying the piecewise linear model, where the model selections are roughly equally split between all the other possibilities although it performs slightly better than the other frameworks when the DGP is an AR(4). In sum, Table 2 shows that weighting the forecast errors has some benefits, albeit modest, since the unweighted fully recursive framework never selects the correct model the highest proportion of times. The 1+log-t weighting scheme works best for the structural models with exogenous variables, while the f-function is best for the piecewise linear model.

Tables 3 to 5 continue, presenting the findings for sample sizes of 50, 100, and 250 respectively. The entries demonstrate the tendency towards asymptotic behaviour, which is again slower for the 2/3-1/3 approach than the others. In general, all of the weighting schemes except 2/3-1/3 are able to identify any data generating processes with good and increasing accuracy. The relative performance of the f- and SE-weighting schemes gradually improves following quite a poor showing for the smallest sample sizes, probably because the divisors in these approaches are themselves data-dependent and therefore noisy and unreliable in such instances. In contrast, the 1+log(t) based function performs well throughout. By 100 observations (Table 4), the correct model identification rate is already well over 80%, and with 250 observations (Table 5), the figure is near 100% in most cases. Hence it is clear that by 100 observations (at least for the DGPs we examine), any forecasting framework will identify the appropriate model with a high degree of dependability, and even 2/3-1/3 has almost caught up. Therefore, the issues that we highlight and the proposal to weight the forecast errors are only relevant in the context of samples smaller than this.

Table 6 moves on to summarise the size and power of the Diebold-Mariano statistic. Throughout, we employ a nominal test size of 10% following the convention in the literature (e.g., Diebold and Mariano, 1995, Tables 1 to 4) and we also employ the asymptotic (standard normal) critical values

throughout. Panel A constructs the size of the test by simulating two separate series of length T (T = 15, 25, 50, 100, or 250) from the same DGP. We do this for each DGP and for each weighting function as described above; for ease of presentation, we average the sizes across the DGPs in the table. The first salient feature of Panel A in Table 6 is that the 2/3-1/3 approach is badly over-sized at small samples, rejecting the null hypothesis 25% of the time for 15 observations and 16% for 25 observations. Perhaps more surprisingly, all of the recursive schemes are reasonably well sized when just 25 observations are available. For the smaller sample sizes (15, 25 or 50 observations), the SE-weighting is marginally preferred, while it becomes slightly undersized for larger T and then the 1+log(t) weighting is closest to the nominal 10% value.

Panel B summarises the power of the test, again assuming a nominal 10% size with asymptotic critical values. Note that in this table we do not make a comparison with the 2/3-1/3 approach due to the over-sizing noted above. The table entries are constructed as follows. We simulate the data according to a specific DGP from the list above, and we then produce forecasts from that model and from the other five. We can then compute five DM statistics – one for each comparison of the correct model accuracy with the wrong model accuracy, for each of the weighting functions. We repeat this analysis by making each of the models the DGP and comparing forecasts from it with those from the other five models and the powers are then averaged across all of the DGPs and comparator models for presentational ease.

The results in Panel B of Table 6 again demonstrate that weighting the forecast errors can lead to some (again modest) improvements, now in terms of correct rejection of the null hypothesis of equal predictive accuracy of competing models, compared with the use of unweighted errors from recursive forecasting. For example, when 50 observations are available, the DM test can correctly distinguish between two models on average 62.69% of the time when fully recursive forecasting is used without weighting, but 64.44% when 1+log(t) weighting is used. The weightings by the standard error of the regression also perform well, and both weighting schemes are uniformly superior to no weighting.

For illustration, Table 7 unpacks the power results that were averaged in Panel B of Table 6, focusing on a sample size of 100 observations.[23] Here, we make each of the six models the DGP and then examine the ability of the DM test with each weighting scheme to distinguish it from each

[23] It would of course be possible to present a similar breakdown for all of the sample sizes that we investigate; however, we do not do so to preserve space.

of the five other models – thus, in total, there are 30 model comparisons. Here, the gain from weighting the forecast errors becomes very evident for some data generating processes. For example, when the DGP is the random walk with drift and the comparator model is an AR(4), the DM test can correctly distinguish them 74.25% of the time when 1+log(t) weights are used but only 62.77% of the time when the squared forecast errors are unweighted. Of all the weighting schemes, the 1+log(t) function provides the highest power most frequently (around one third of the time) with f- and regression standard error weighting yielding the highest power for around a quarter of the DGP-comparator model combinations. For the individual comparisons, the no-weighting approach is never the best among the various schemes although in some instances it comes fairly close.

4. Forecast Evaluation with Real Data

4.1 Background

In this section, we now turn our attention to forecast evaluation with actual rather than simulated data based on the various weighting schemes described above. Although there is an almost infinite number of possible series and sample periods that we could have investigated, we elect to employ the data used by Meese and Rogoff (1983) given the prominence of their conclusions within the literature and the extent to which their work has become regarded as a defining moment in the forecasting world. In essence, their finding that the simple random walk model could produce better forecasts than a raft of more sophisticated time-series and structural models, and also better than the forward rate obtained from the financial markets, has become a stylised fact.

An extremely large body of research has since tried to further test and explain these findings. Cheung, Chinn and Pascual (2005) show that the key conclusions still hold when an additional two decades of more recent data are added to the sample. That the random walk model could not be beaten has been viewed in the literature as somewhat of a puzzle. Meese and Rogoff advocate out-of-sample testing as a way to circumvent possible endogeneity issues that may bias in-sample tests of model adequacy, but if such bias exists in the model, it is surely bound to adversely affect forecast performance as well. They suggest that the poor performance of the macroeconomic models, which are built on apparently strong theoretical foundations, may be attributable to simultaneous equations bias, parameter instability or another form of model misspecification. Bacchetta, van Wincoop and Beutler (2009) comprehensively test but rule out parameter instability as an explanation of Meese and Rogoff's findings and thus the search for a reliable justification continues. However, it is perhaps useful to state at the outset that we do not intend to contribute to the debate about why the random walk model appears to describe exchange rate movements so

well; rather, we now revisit their conclusions using the same data and forecasting models but using the various squared error weighting schemes highlighted in Section 2.[24]

4.2 Data and Methodology

Equation (1) in Meese and Rogoff (1983) summarises the structural models they implement (with time subscripts suppressed):

$$s = a_0 + a_1\left(m - \dot{m}\right) + a_2\left(y - \dot{y}\right) + a_3\left(r_s - \dot{r}_s\right) + a_4\left(\pi^e - \dot{\pi}^e\right) + a_5 TB + a_6 T\dot{B} + u \qquad (29)$$

where s is the log of the exchange rate; m, y, r_s, π^e and TB are domestic macro variables, while $\dot{m}$, $\dot{y}$, $\dot{r}_s$, $\dot{\pi}^e$ and $T\dot{B}$ are their foreign counterparts; m is the log of the money supply, y is the real income, r_s is the short term interest rate, π^e is the expected long run inflation, TB is the trade balance; finally, u is a disturbance term. For $a_4 = a_5 = a_6 = 0$ we obtain the Frenkel-Bilson model, while with $a_5 = a_6 = 0$ we get the Dornbusch-Frenkel model. In their setup, the domestic economy is the United States and the foreign economies are Germany, Japan and UK. The unrestricted model in (29) is what Meese and Rogoff (1983) call the Hooper-Morton model. Meese and Rogoff (1983) subsequently estimate six univariate time series models: four AR models, with lags determined, respectively, using the Akaike criterion, Schwartz criterion and a deterministic rule – lag length = sample size/ log(sample size), the latter using both the original and an exponentially weighted sample, and the random walk, with or without a drift. They also estimate a multivariate time series model, a VAR model where each variable (in their structural model (1)) is regressed against its own lagged values as well as lagged values of the other variables, the lag length being estimated using Parzen's (1975) selection criterion. We shall repeat the exercise above by simulating one time series model, what they call the 'long AR', where the lag length is given by a deterministic formula depending on sample size, and evaluate the forecasts obtained with the structural and the other time series models, using the same criteria as above.

Meese and Rogoff's sample covers the months January 1973 to December 1982, although some of the data are missing for certain variables and thus the common period for all the series July 1973 to July 1982 is used throughout. In our recursive framework, the initial model estimation runs to July 1974 with the first one-step ahead forecast being made for August 1974. Then the estimation window is extended by one month at a time, the models re-estimated and a new set of one-step

[24] We note, of course, that their sample is more than two decades old. However, we do not consider this an issue since our interest is in re-evaluating the models chosen when the new weighting functions are employed and as such, the sample period is irrelevant.

ahead forecasts produced until the sample is exhausted. This results in a total of 100 out of sample forecasts, and the forecast errors are squared and weighted using the six approaches (equally weighted (fully recursive and 2/3-1/3 split), *t*-weighted, 1/*f*-weighted, 1/SE(weighted), and 1+log(*t*)-weighted). The 2/3-1/3 approach retains 36 observations (August 1979-July 1982) as the hold-out sample. We follow Meese and Rogoff in using the actual values of any exogenous variables required in the models despite the fact that they were not observable at that time and in reality would have needed to have been forecast. As they note, this provides a considerable artificial advantage to such models which makes their poor performance all the more surprising.

4.3 Empirical Results

When the Meese-Rogoff data are employed with the various weighting schemes, we find for all three currency pairs that the random walk model has the lowest MSE when either weighted or unweighted squared errors are employed, or if the sample is split on a 2/3-1/3 basis. Almost invariably, the random walk with drift has the second lowest MSE, the forward rate is third and the AR(12) is fourth with the structural models all performing worst. Thus the split in performance accuracy between the time-series and structural models is so stark that use of different weighting functions will not overturn the well-established conclusions. We therefore do not tabulate these findings on the MSE figures themselves or the model rankings to preserve space but instead the results presented are for the Diebold-Mariano test statistics. Panels A to C of Table 8 are for British pound – US dollar, the German mark – dollar, and the Japanese yen – dollar exchange rates respectively and *p*-values from two-sided asymptotic critical values are given in parentheses. Each test represents a comparison with the random walk model since this is arguably the most relevant base model.

Echoing the findings from the Monte Carlo we conducted above, the real data application demonstrate that using a 2/3-1/3 approach is inadvisable. For the GBP_USD rate (Panel A), the DM statistics are significant for any weighting schemes based on the use of all available data for prediction (i.e. commencing the forecasts at observation $k_{max}+1$) but not for 2/3-1/3 and this is probably a function of the lack of power resulting from the reduced effective sample in the latter case. Similarly, in the case of the DEM_USD (Panel B), 2/3-1/3 leads to a rejection of the null of equal predictive capacity when comparing the random walk with the random walk with drift, or comparing the random walk with the AR(12) but for none of the weighting schemes. Given the Monte Carlo results, this appears likely to be a spurious rejection of the null due to the severe over-sizing of the former approach.

More interestingly, the no-weighting fully recursive approach shows no statistically significant DM statistics when comparing the random walk with the forward rate for the GBP_USD, and yet all of the weighting schemes do. Similarly, the *t*- and *f*-weighting schemes show rejections at the 10% level or better for the VAR(2) but the unweighted (and other weighting functions) squared errors do not. This is suggestive that weighting the squared forecast errors may lead to qualitatively different conclusions in some cases, and we conjecture that the variation would be even greater if the total sample size were smaller.

5. Conclusions

This paper has proposed and employed an approach to determining the most accurate forecaster among a set of competing models when a limited sample of data is available. Relevant applications could include studies of emerging markets, or those which use data that is only available at relatively low frequency – such as, for example, macroeconomic data or data on new or less frequently traded assets (e.g., real estate). The core of our idea is essentially to weight the forecast errors according to the amount of information contained in the (in-sample) model estimation. We propose a weighting scheme based on the ex ante forecast uncertainty but also a family of simpler approximations to this.

The approach that we advocate, where recursive forecasting begins at the earliest possible opportunity, can be seen as a viable alternative to pure in-sample testing but where the power of out-of-sample tests of forecast accuracy is not reduced by 'losing' a large proportion of the data for initial model estimation.[25] We demonstrate that in small samples there are significant improvements brought about by employing any member of this family rather than the standard (2/3-1/3 sample split) approach. Firstly, we show that the proposed approaches deliver the correct model a considerably higher proportion of the time than the standard approach in the literature, which we stylise as using two thirds of the data for initial in-sample model estimation and to equally weight the forecast errors for the remaining third. Our simulation study also shows that they deliver Diebold Mariano (DM) tests which are much better sized than in the case of the standard approach and that the proposed weighting schemes result in an improvement (at times considerable) over the unweighted fully recursive scheme.

[25] As we also highlight later in the paper, there is nevertheless a correspondence in some circumstances between one of the weighted recursive approaches that we use and an established in-sample, goodness of fit criterion (RSS).

We then apply the new criteria to the exchange rate data employed in the classic Meese and Rogoff (1983) study. In common with the results for the simulated data, we find that there are relevant differences between the recursive weighting schemes that we propose and the standard 2/3-1/3 sample split approach. When we use the Diebold Mariano methodology in this context, we find that the weighted recursive schemes and the standard approach can lead to different conclusions regarding the significance of the differences in accuracy between competing forecasts. Combining this finding with the fact that our Monte Carlo study on the size of the Diebold Mariano test showed that the standard approach can result in significant over-sizing, we conclude that the 2/3-1/3 approach can erroneously signal differences in forecast accuracy when there is none, whereas the proposed weighting approaches, which appeared to produce much better sized DM tests, would correctly identify two models as having equal forecast accuracy.

We conclude by noting that while an in-sample comparative evaluation of models is more powerful than pseudo-out of sample-based tests under fairly weak conditions, if researchers continue to express a preference for the latter, the current practice of sample-splitting is unadvisable and a fully recursive approach exhibits better properties, especially in small samples. Hence the implications of our findings for applied researchers are very clear and can be summarised in two statements. First, it is considerably better to employ as much of the available data as possible in the evaluation of out-of-sample forecasts than to waste a large slab of information in estimating the models in-sample. Second, given that the forecasting begins recursively as early as possible in the sample, weighting the forecasts will further improve the effectiveness of the identification of the correct model, but only slightly. Weighting can also improve the size of the Diebold Mariano test for equal predictive accuracy, albeit not to a great extent; improvements in the power of such tests, however, can be significant in certain cases. If weighting is undertaken, this can be just as well achieved by using a heuristic function based on the sample size (either linear or logarithmic) as by using a more sophisticated approach based on in-sample estimation accuracy or ex ante one-step ahead forecasting accuracy.

There are many natural ways in which the analysis conducted in this paper could be extended, in part as a result of the numerous simplifications we made in the research design to aid tractability. First, we have focused specifically on one-step ahead forecasts and it would be of considerable interest to consider whether our results also apply when longer term forecasts are required. Second, our approach has considered only linear models while empirical research, especially in finance and related fields, is frequently concerned with non-linear models – for example, when forecasting

volatilities or correlations. Third, we have assumed that mean squared errors represent the model evaluation criterion of interest but it would be possible to instead adopt mean absolute errors or economically motivated error aggregation metrics. Finally, it would be of interest to more fully investigate the impact of structural breaks or more general periods of parameter instability in the data generating process on the effectiveness of the various forecast evaluation measures.

References

Ashley, R. (2003) Statistically Significant forecasting improvements: how much out-of-sample data is likely necessary? International Journal of Forecasting 19, 229-239.

Ashley, R., Granger, C.W.J., and Schmalensee, R. (1980) Advertising and aggregate consumption: an analysis of causality. Econometrica 48, 1149-1168.

Bacchetta, P., van Wincoop, E. & Beutler, T. (2009) Can parameter instability explain the Meese-Rogoff puzzle? CEPR Discussion Paper DP7383.

Brown, R.L., Durbin, J. & Evans, J.M. (1975) Techniques for testing the constancy of regression relationships over time. Journal of the Royal Statistical Society, Series B (Methodological) 37(2), 149-192.

Busetti, F. & Marcucci, J. (2013) Comparing forecast accuracy: a Monte Carlo investigation. International Journal of Forecasting 29, 13-27.

Cheung, Y-W., Chinn, M.D. & Pascual, A.G. (2003) Empirical exchange rate models of the nineties: are any fit to survive? Journal of International Money and Finance. 24, 1150-1175.

Clark, T.E. & McCracken, M.W. (2001) Tests of equal forecast accuracy and encompassing for nested models. Journal of Econometrics 105, 85-110.

Clark, T.E. & McCracken, M.W. (2005a) Evaluating direct multi-step forecasts. Econometric Reviews 24, 369-404.

Clark, T.E. & McCracken, M.W. (2005b) The power of tests of predictive ability in the presence of structural breaks. Journal of Econometrics 124, 1-31.

Clark, T.E. & West, K.D. (2006) Using out-of-sample mean squared prediction errors to test the Martingale difference hypothesis. Journal of Econometrics 135, 155-186.

Clark, T.E. & West, K.D. (2007) Approximately normal tests for equal predictive accuracy in nested models. Journal of Econometrics 138, 291-311.

Clements, M.P. (2005) Evaluating econometric forecasts of economic and financial variables Basingstoke: Palgrave Macmillan.

Clements, M.P. & Hendry, D.F. (1993) On the limitations of comparing mean squared forecast errors. Journal of Forecasting 12, 617-676.

Diebold, F.X. & Mariano, R.S. (1995) Comparing predictive accuracy. Journal of Business and Economic Statistics 13(3), 253-263.

Faust, J., Rogers, J.H. & Wright, J.H. (2004) News and noise in G-7 GDP announcements. Journal of Money, Credit and Banking 37(3), 403-19.

Giacomini, R. & Rossi, B. (2010) Forecast comparisons in unstable environments. Journal of Applied Econometrics 25, 595-620.

Hansen. P.R. & Timmerman, A. (2012) Choice of sample split in out-of-sample forecast evaluation. Working Paper, European University Institute.

Hansen, P.R. & Timmermann, A. (2013) Equivalence between out-of-sample forecast comparisons and Wald statistics. Working Paper, European University Institute and University of California, San Diego.

Harvey, A.C. (1990): The Econometric Analysis of Time Series, 2nd Edition. Oxford: Philip Allan.

Harvey, D.I., Leybourne, S. & Newbold, P. (1997) Testing the equality of prediction mean squared errors. International Journal of Forecasting 13, 281-291.

Harvey, D.I., Leybourne, S. & Newbold, P. (1998) Tests of forecast encompassing. Journal of Business and Economic Statistics 16, 254-259.

Inoue, A. & Kilian, L. (2005) In-sample or out-of-Sample tests of predictability – which one should we use? Econometric Reviews 23(4), 371-402.

Inoue, A. & Kilian, L. (2006) On the selection of forecasting models. Journal of Econometrics 130(2), 273-306.

McCracken, M.W. (2000) Robust out-of-sample inference. Journal of Econometrics 99, 195-223.

McCracken, M.W. (2007) Asymptotics for out-of-sample tests of Granger causality. Journal of Econometrics 140, 719-752.
Meese, R. A. & Rogoff, K. (1983) Empirical exchange rate models of the seventies: do they fit out of sample? Journal of International Economics 14, 3-24.
Parzen, E. (1975) Multiple time series: determining the order of approximating autoregressive schemes. Technical Report No. 23, State University of New York at Buffalo.
Philips, G.D.A. & Harvey, A.C. (1974) A simple test for serial correlation in regression analysis. Journal of the American Statistical Association 69(348), 935-939.
Rossi, B. & Inoue, A. (2012) Out-of-sample forecast tests robust to the window size choice Working Paper, Duke University.
West, K.D. (1996) Asymptotic inference about predictive ability. Econometrica 64(5), 1067-1084..
West, K.D. (2006) Forecast evaluation. In Elliott, G., Granger, C.W.J. & Timmerman, A. (eds.) Handbook of Economic Forecasting, Volume 1, North Holland: Elsevier.
West, K.D. & McCracken, M.W. (1998) Regression-based tests of predictive ability. International Economic Review 39(4), 817-840.

Figure 1: Illustration of the Various Weighting Schemes for *T*=30

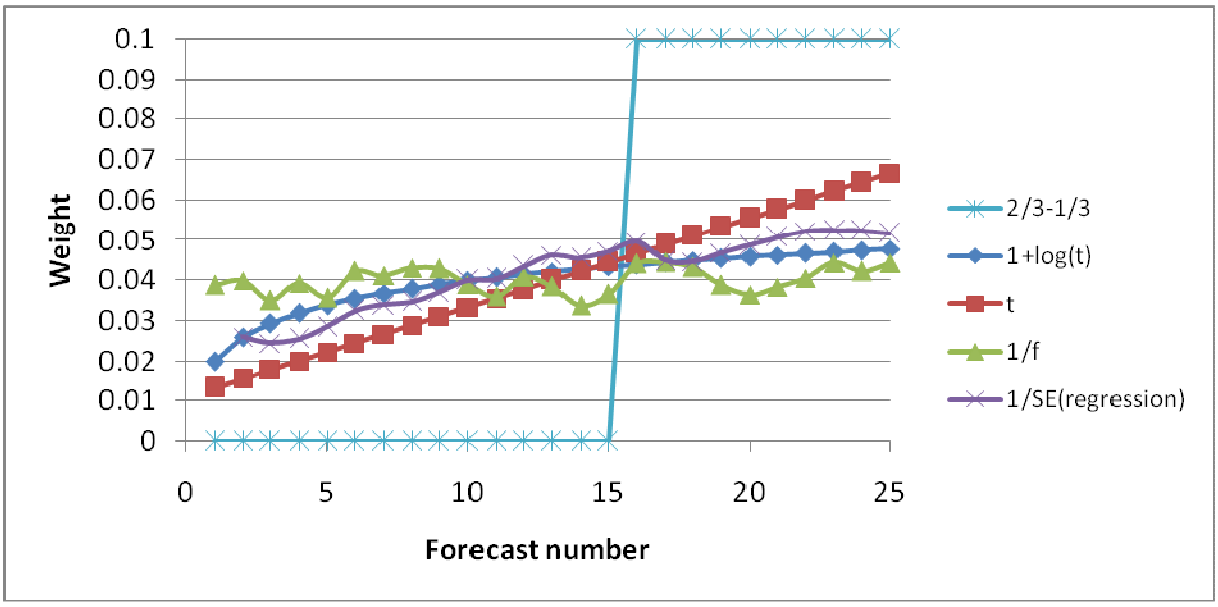

Table 1: Monte Carlo Results for *T*=15

		Panel A: Correct Model is a Random Walk with Drift				
	RWD	AR(4)	Linear Structural	Quadratic Structural	Multiplicative Structural	Piecewise linear
Recursive from k+1 (no weighting)	94.0	1.3	0.9	0.8	0.7	2.3
Recursive with t-weighting	94.0	1.3	0.9	0.8	0.8	2.3
Recursive with 1/f-weighting	94.5	0.9	0.8	0.8	0.7	2.2
Recursive with 1/SE(reg) weighting	88.7	4.2	1.2	1.3	1.2	3.5
Recursive with (1+log(t)) weighting	93.5	1.7	0.8	0.9	0.8	2.3
2/3 - 1/3 rule (no weighting)	52.4	28.4	2.1	2.3	2.3	12.5
		Panel B: Correct model is an AR(4)				
	RWD	AR(4)	Linear Structural	Quadratic Structural	Multiplicative Structural	Piecewise linear
Recursive from k+1 (no weighting)	25.6	73.4	0.2	0.2	0.2	0.4
Recursive with t-weighting	25.1	73.9	0.2	0.2	0.2	0.4
Recursive with 1/f-weighting	26.8	72.2	0.2	0.2	0.2	0.4
Recursive with 1/SE(reg) weighting	12.4	85.6	0.3	0.4	0.4	0.9
Recursive with (1+log(t)) weighting	19.7	79.3	0.2	0.2	0.2	0.4
2/3 - 1/3 rule (no weighting)	18.7	61.7	4.9	4.9	5.0	4.8
		Panel C: Correct Model is a Linear Structural				
	RWD	AR(4)	Linear Structural	Quadratic Structural	Multiplicative Structural	Piecewise linear
Recursive from k+1 (no weighting)	1.3	0.7	70.7	6.8	6.6	13.9
Recursive with t-weighting	1.3	0.8	70.7	6.8	6.6	13.8
Recursive with 1/f-weighting	2.7	0.8	68.2	7.4	6.6	14.3
Recursive with 1/SE(reg) weighting	2.0	1.9	65.1	8.7	7.8	14.6
Recursive with (1+log(t)) weighting	1.4	0.8	71.2	6.9	6.4	13.4
2/3 - 1/3 rule (no weighting)	16.6	15.9	31.6	12.5	10.6	12.8
		Panel D: Correct Model is a Quadratic Structural				
	RWD	AR(4)	Linear Structural	Quadratic Structural	Multiplicative Structural	Piecewise linear
Recursive from k+1 (no weighting)	1.2	0.9	6.9	69.7	7.4	13.9
Recursive with t-weighting	1.2	0.9	6.8	69.8	7.3	14.0
Recursive with 1/f-weighting	2.4	1.0	7.6	67.5	7.7	13.8
Recursive with 1/SE(reg) weighting	1.8	1.8	8.2	66.7	8.6	12.9
Recursive with (1+log(t)) weighting	1.3	1.1	6.8	70.8	7.4	12.7
2/3 - 1/3 rule (no weighting)	15.3	15.2	12.9	32.9	11.7	12.0
		Panel E: Correct Model is an Multiplicative Structural				
	RWD	AR(4)	Linear Structural	Quadratic Structural	Multiplicative Structural	Piecewise linear
Recursive from k+1 (no weighting)	1.5	0.9	8.2	8.5	61.9	19.0
Recursive with t-weighting	1.5	0.9	8.2	8.5	61.9	19.0
Recursive with 1/f-weighting	3.2	0.8	8.5	8.7	59.9	18.9
Recursive with 1/SE(reg) weighting	2.6	1.9	9.7	10.3	57.8	17.8
Recursive with (1+log(t)) weighting	1.7	1.0	8.2	8.7	62.2	18.3
2/3 - 1/3 rule (no weighting)	17.5	16.7	12.9	13.0	26.9	13.2
		Panel F: Correct Model is a Piecewise Linear				
	RWD	AR(4)	Linear Structural	Quadratic Structural	Multiplicative Structural	Piecewise linear
Recursive from k+1 (no weighting)	6.9	0.9	13.5	12.5	14.1	52.2
Recursive with t-weighting	6.9	1.0	13.6	12.5	14.1	52.1
Recursive with 1/f-weighting	11.9	1.0	12.5	12.9	13.0	48.8
Recursive with 1/SE(reg) weighting	9.7	2.1	14.2	14.2	14.5	45.3
Recursive with (1+log(t)) weighting	7.3	1.0	14.1	13.0	14.5	50.2
2/3 - 1/3 rule (no weighting)	21.6	17.6	13.1	12.6	12.8	22.4

Table 2: Monte Carlo Results for *T*=25

		Panel A: Correct Model is a Random Walk with Drift				
	RWD	AR(4)	Linear Structural	Quadratic Structural	Multiplicative Structural	Piecewise linear
Recursive from k+1 (no weighting)	99.0	0.6	0.0	0.1	0.0	0.2
Recursive with t-weighting	99.0	0.6	0.0	0.1	0.0	0.2
Recursive with 1/f-weighting	99.3	0.3	0.0	0.1	0.0	0.2
Recursive with 1/SE(reg) weighting	98.0	1.4	0.1	0.1	0.1	0.4
Recursive with (1+log(t)) weighting	98.9	0.8	0.0	0.1	0.0	0.2
2/3 - 1/3 rule (no weighting)	70.1	22.8	0.4	0.4	0.4	5.9
		Panel B: Correct model is an AR(4)				
	RWD	AR(4)	Linear Structural	Quadratic Structural	Multiplicative Structural	Piecewise linear
Recursive from k+1 (no weighting)	18.4	81.4	0.0	0.0	0.1	0.1
Recursive with t-weighting	17.9	81.9	0.0	0.0	0.1	0.1
Recursive with 1/f-weighting	18.2	81.7	0.0	0.0	0.0	0.1
Recursive with 1/SE(reg) weighting	9.5	90.1	0.0	0.0	0.1	0.2
Recursive with (1+log(t)) weighting	12.3	87.5	0.0	0.0	0.0	0.1
2/3 - 1/3 rule (no weighting)	2.5	95.9	0.3	0.4	0.4	0.4
		Panel C: Correct Model is a Linear Structural				
	RWD	AR(4)	Linear Structural	Quadratic Structural	Multiplicative Structural	Piecewise linear
Recursive from k+1 (no weighting)	0.1	0.2	88.2	2.7	2.2	6.7
Recursive with t-weighting	0.1	0.2	88.4	2.6	2.2	6.5
Recursive with 1/f-weighting	0.3	0.1	88.3	2.7	2.2	6.4
Recursive with 1/SE(reg) weighting	0.1	0.3	85.3	3.7	3.0	7.6
Recursive with (1+log(t)) weighting	0.1	0.2	89.2	2.6	2.2	5.6
2/3 - 1/3 rule (no weighting)	3.6	8.2	63.8	9.1	7.8	7.5
		Panel D: Correct Model is a Quadratic Structural				
	RWD	AR(4)	Linear Structural	Quadratic Structural	Multiplicative Structural	Piecewise linear
Recursive from k+1 (no weighting)	0.1	0.2	2.6	86.4	3.1	7.7
Recursive with t-weighting	0.1	0.2	2.5	86.7	3.0	7.5
Recursive with 1/f-weighting	0.2	0.2	2.8	86.9	3.2	6.7
Recursive with 1/SE(reg) weighting	0.1	0.4	3.0	86.1	3.4	7.1
Recursive with (1+log(t)) weighting	0.1	0.2	2.2	88.5	2.7	6.3
2/3 - 1/3 rule (no weighting)	3.1	6.6	8.0	67.7	8.2	6.5
		Panel E: Correct Model is an Multiplicative Structural				
	RWD	AR(4)	Linear Structural	Quadratic Structural	Multiplicative Structural	Piecewise linear
Recursive from k+1 (no weighting)	0.1	0.3	3.6	4.7	79.8	11.6
Recursive with t-weighting	0.1	0.3	3.5	4.6	80.2	11.3
Recursive with 1/f-weighting	0.4	0.2	3.6	4.4	80.8	10.6
Recursive with 1/SE(reg) weighting	0.2	0.5	4.6	5.5	78.6	10.7
Recursive with (1+log(t)) weighting	0.2	0.4	3.6	4.2	82.1	9.6
2/3 - 1/3 rule (no weighting)	3.8	9.9	10.0	9.9	58.1	8.4
		Panel F: Correct Model is a Piecewise Linear				
	RWD	AR(4)	Linear Structural	Quadratic Structural	Multiplicative Structural	Piecewise linear
Recursive from k+1 (no weighting)	1.5	0.4	10.8	9.9	10.8	66.6
Recursive with t-weighting	1.5	0.4	10.9	9.9	11.0	66.3
Recursive with 1/f-weighting	3.4	0.5	10.0	9.5	9.9	66.7
Recursive with 1/SE(reg) weighting	1.9	1.0	11.3	10.8	11.5	63.5
Recursive with (1+log(t)) weighting	1.9	0.6	11.8	10.6	11.7	63.3
2/3 - 1/3 rule (no weighting)	7.4	13.5	14.5	13.2	14.0	37.6

Table 3: Monte Carlo Results for *T*=50

		Panel A: Correct Model is a Random Walk with Drift				
	RWD	AR(4)	Linear Structural	Quadratic Structural	Multiplicative Structural	Piecewise linear
Recursive from k+1 (no weighting)	99.9	0.1	0.0	0.0	0.0	0.0
Recursive with t-weighting	99.9	0.1	0.0	0.0	0.0	0.0
Recursive with 1/f-weighting	99.9	0.1	0.0	0.0	0.0	0.0
Recursive with 1/SE(reg) weighting	99.8	0.2	0.0	0.0	0.0	0.0
Recursive with (1+log(t)) weighting	99.8	0.2	0.0	0.0	0.0	0.0
2/3 - 1/3 rule (no weighting)	80.3	19.1	0.0	0.0	0.0	0.5
		Panel B: Correct model is an AR(4)				
	RWD	AR(4)	Linear Structural	Quadratic Structural	Multiplicative Structural	Piecewise linear
Recursive from k+1 (no weighting)	9.9	90.1	0.0	0.0	0.0	0.0
Recursive with t-weighting	9.4	90.6	0.0	0.0	0.0	0.0
Recursive with 1/f-weighting	8.3	91.7	0.0	0.0	0.0	0.0
Recursive with 1/SE(reg) weighting	4.4	95.6	0.0	0.0	0.0	0.0
Recursive with (1+log(t)) weighting	4.9	95.1	0.0	0.0	0.0	0.0
2/3 - 1/3 rule (no weighting)	0.7	98.9	0.1	0.1	0.1	0.1
		Panel C: Correct Model is a Linear Structural				
	RWD	AR(4)	Linear Structural	Quadratic Structural	Multiplicative Structural	Piecewise linear
Recursive from k+1 (no weighting)	0.0	0.0	98.9	0.3	0.2	0.6
Recursive with t-weighting	0.0	0.0	99.1	0.2	0.2	0.5
Recursive with 1/f-weighting	0.0	0.0	99.2	0.2	0.1	0.5
Recursive with 1/SE(reg) weighting	0.0	0.0	98.4	0.4	0.3	1.0
Recursive with (1+log(t)) weighting	0.0	0.0	99.5	0.2	0.1	0.3
2/3 - 1/3 rule (no weighting)	0.4	3.1	85.8	3.9	3.7	3.0
		Panel D: Correct Model is a Quadratic Structural				
	RWD	AR(4)	Linear Structural	Quadratic Structural	Multiplicative Structural	Piecewise linear
Recursive from k+1 (no weighting)	0.0	0.0	0.4	96.6	0.7	2.3
Recursive with t-weighting	0.0	0.0	0.4	96.8	0.7	2.1
Recursive with 1/f-weighting	0.0	0.0	0.4	97.6	0.6	1.3
Recursive with 1/SE(reg) weighting	0.0	0.1	0.3	97.3	0.7	1.6
Recursive with (1+log(t)) weighting	0.0	0.0	0.2	98.2	0.5	1.1
2/3 - 1/3 rule (no weighting)	0.2	2.5	2.5	88.9	3.5	2.4
		Panel E: Correct Model is an Multiplicative Structural				
	RWD	AR(4)	Linear Structural	Quadratic Structural	Multiplicative Structural	Piecewise linear
Recursive from k+1 (no weighting)	0.0	0.0	0.6	1.1	95.3	3.0
Recursive with t-weighting	0.0	0.0	0.6	1.0	95.6	2.8
Recursive with 1/f-weighting	0.0	0.0	0.4	0.7	97.0	1.9
Recursive with 1/SE(reg) weighting	0.0	0.1	0.7	1.0	95.4	2.8
Recursive with (1+log(t)) weighting	0.0	0.0	0.4	0.7	97.3	1.7
2/3 - 1/3 rule (no weighting)	0.6	4.3	4.0	5.4	81.9	3.7
		Panel F: Correct Model is a Piecewise Linear				
	RWD	AR(4)	Linear Structural	Quadratic Structural	Multiplicative Structural	Piecewise linear
Recursive from k+1 (no weighting)	0.0	0.1	5.9	5.4	5.4	83.3
Recursive with t-weighting	0.0	0.1	6.0	5.4	5.4	83.1
Recursive with 1/f-weighting	0.1	0.1	5.4	4.9	4.9	84.6
Recursive with 1/SE(reg) weighting	0.1	0.3	6.1	5.6	5.9	82.0
Recursive with (1+log(t)) weighting	0.1	0.3	6.8	5.9	6.1	81.0
2/3 - 1/3 rule (no weighting)	2.0	9.8	9.7	10.7	9.9	57.8

Table 4: Monte Carlo Results for *T*=100

		Panel A: Correct Model is a Random Walk with Drift				
	RWD	AR(4)	Linear Structural	Quadratic Structural	Multiplicative Structural	Piecewise linear
Recursive from k+1 (no weighting)	100.0	0.0	0.0	0.0	0.0	0.0
Recursive with t-weighting	100.0	0.0	0.0	0.0	0.0	0.0
Recursive with 1/f-weighting	100.0	0.0	0.0	0.0	0.0	0.0
Recursive with 1/SE(reg) weighting	100.0	0.0	0.0	0.0	0.0	0.0
Recursive with (1+log(t)) weighting	100.0	0.0	0.0	0.0	0.0	0.0
2/3 - 1/3 rule (no weighting)	82.8	17.2	0.0	0.0	0.0	0.0
		Panel B: Correct model is an AR(4)				
	RWD	AR(4)	Linear Structural	Quadratic Structural	Multiplicative Structural	Piecewise linear
Recursive from k+1 (no weighting)	5.3	94.7	0.0	0.0	0.0	0.0
Recursive with t-weighting	5.0	95.0	0.0	0.0	0.0	0.0
Recursive with 1/f-weighting	4.1	95.9	0.0	0.0	0.0	0.0
Recursive with 1/SE(reg) weighting	1.7	98.3	0.0	0.0	0.0	0.0
Recursive with (1+log(t)) weighting	2.4	97.6	0.0	0.0	0.0	0.0
2/3 - 1/3 rule (no weighting)	0.1	99.9	0.0	0.0	0.0	0.0
		Panel C: Correct Model is a Linear Structural				
	RWD	AR(4)	Linear Structural	Quadratic Structural	Multiplicative Structural	Piecewise linear
Recursive from k+1 (no weighting)	0.0	0.0	100.0	0.0	0.0	0.0
Recursive with t-weighting	0.0	0.0	100.0	0.0	0.0	0.0
Recursive with 1/f-weighting	0.0	0.0	100.0	0.0	0.0	0.0
Recursive with 1/SE(reg) weighting	0.0	0.0	100.0	0.0	0.0	0.0
Recursive with (1+log(t)) weighting	0.0	0.0	100.0	0.0	0.0	0.0
2/3 - 1/3 rule (no weighting)	0.0	0.6	97.2	0.9	0.8	0.5
		Panel D: Correct Model is a Quadratic Structural				
	RWD	AR(4)	Linear Structural	Quadratic Structural	Multiplicative Structural	Piecewise linear
Recursive from k+1 (no weighting)	0.0	0.0	0.1	98.7	0.3	0.9
Recursive with t-weighting	0.0	0.0	0.1	98.9	0.3	0.7
Recursive with 1/f-weighting	0.0	0.0	0.1	99.3	0.2	0.4
Recursive with 1/SE(reg) weighting	0.0	0.0	0.1	99.3	0.2	0.5
Recursive with (1+log(t)) weighting	0.0	0.0	0.0	99.7	0.1	0.2
2/3 - 1/3 rule (no weighting)	0.0	0.4	0.3	98.2	0.6	0.5
		Panel E: Correct Model is an Multiplicative Structural				
	RWD	AR(4)	Linear Structural	Quadratic Structural	Multiplicative Structural	Piecewise linear
Recursive from k+1 (no weighting)	0.0	0.0	0.1	0.3	99.1	0.6
Recursive with t-weighting	0.0	0.0	0.1	0.3	99.3	0.5
Recursive with 1/f-weighting	0.0	0.0	0.0	0.2	99.6	0.2
Recursive with 1/SE(reg) weighting	0.0	0.0	0.1	0.3	99.3	0.4
Recursive with (1+log(t)) weighting	0.0	0.0	0.0	0.1	99.8	0.1
2/3 - 1/3 rule (no weighting)	0.0	0.9	0.9	1.4	95.7	1.0
		Panel F: Correct Model is a Piecewise Linear				
	RWD	AR(4)	Linear Structural	Quadratic Structural	Multiplicative Structural	Piecewise linear
Recursive from k+1 (no weighting)	0.0	0.1	1.9	1.6	1.5	95.0
Recursive with t-weighting	0.0	0.1	1.9	1.6	1.5	94.9
Recursive with 1/f-weighting	0.0	0.0	1.7	1.5	1.2	95.5
Recursive with 1/SE(reg) weighting	0.0	0.1	1.9	1.8	1.6	94.6
Recursive with (1+log(t)) weighting	0.0	0.1	2.2	1.8	1.9	94.0
2/3 - 1/3 rule (no weighting)	0.2	6.2	5.3	5.7	4.8	77.9

Table 5: Monte Carlo Results for *T*=250

		Panel A: Correct Model is a Random Walk with Drift				
	RWD	AR(4)	Linear Structural	Quadratic Structural	Multiplicative Structural	Piecewise linear
Recursive from k+1 (no weighting)	100.0	0.0	0.0	0.0	0.0	0.0
Recursive with t-weighting	100.0	0.0	0.0	0.0	0.0	0.0
Recursive with 1/f-weighting	100.0	0.0	0.0	0.0	0.0	0.0
Recursive with 1/SE(reg) weighting	100.0	0.0	0.0	0.0	0.0	0.0
Recursive with (1+log(t)) weighting	100.0	0.0	0.0	0.0	0.0	0.0
2/3 - 1/3 rule (no weighting)	84.5	15.5	0.0	0.0	0.0	0.0
		Panel B: Correct model is an AR(4)				
	RWD	AR(4)	Linear Structural	Quadratic Structural	Multiplicative Structural	Piecewise linear
Recursive from k+1 (no weighting)	2.6	97.4	0.0	0.0	0.0	0.0
Recursive with t-weighting	2.4	97.6	0.0	0.0	0.0	0.0
Recursive with 1/f-weighting	1.9	98.1	0.0	0.0	0.0	0.0
Recursive with 1/SE(reg) weighting	0.5	99.5	0.0	0.0	0.0	0.0
Recursive with (1+log(t)) weighting	1.2	98.8	0.0	0.0	0.0	0.0
2/3 - 1/3 rule (no weighting)	0.0	100.0	0.0	0.0	0.0	0.0
		Panel C: Correct Model is a Linear Structural				
	RWD	AR(4)	Linear Structural	Quadratic Structural	Multiplicative Structural	Piecewise linear
Recursive from k+1 (no weighting)	0.0	0.0	100.0	0.0	0.0	0.0
Recursive with t-weighting	0.0	0.0	100.0	0.0	0.0	0.0
Recursive with 1/f-weighting	0.0	0.0	100.0	0.0	0.0	0.0
Recursive with 1/SE(reg) weighting	0.0	0.0	100.0	0.0	0.0	0.0
Recursive with (1+log(t)) weighting	0.0	0.0	100.0	0.0	0.0	0.0
2/3 - 1/3 rule (no weighting)	0.0	0.0	100.0	0.0	0.0	0.0
		Panel D: Correct Model is a Quadratic Structural				
	RWD	AR(4)	Linear Structural	Quadratic Structural	Multiplicative Structural	Piecewise linear
Recursive from k+1 (no weighting)	0.0	0.0	0.0	99.6	0.2	0.2
Recursive with t-weighting	0.0	0.0	0.0	99.7	0.2	0.2
Recursive with 1/f-weighting	0.0	0.0	0.0	99.8	0.1	0.1
Recursive with 1/SE(reg) weighting	0.0	0.0	0.0	99.8	0.1	0.1
Recursive with (1+log(t)) weighting	0.0	0.0	0.0	99.9	0.1	0.0
2/3 - 1/3 rule (no weighting)	0.0	0.0	0.0	100.0	0.0	0.0
		Panel E: Correct Model is an Multiplicative Structural				
	RWD	AR(4)	Linear Structural	Quadratic Structural	Multiplicative Structural	Piecewise linear
Recursive from k+1 (no weighting)	0.0	0.0	0.0	0.1	99.8	0.1
Recursive with t-weighting	0.0	0.0	0.0	0.1	99.8	0.1
Recursive with 1/f-weighting	0.0	0.0	0.0	0.0	99.9	0.0
Recursive with 1/SE(reg) weighting	0.0	0.0	0.0	0.0	99.9	0.1
Recursive with (1+log(t)) weighting	0.0	0.0	0.0	0.0	100.0	0.0
2/3 - 1/3 rule (no weighting)	0.0	0.0	0.0	0.1	99.9	0.0
		Panel F: Correct Model is a Piecewise Linear				
	RWD	AR(4)	Linear Structural	Quadratic Structural	Multiplicative Structural	Piecewise linear
Recursive from k+1 (no weighting)	0.0	0.0	0.0	0.0	0.0	99.9
Recursive with t-weighting	0.0	0.0	0.0	0.0	0.0	99.9
Recursive with 1/f-weighting	0.0	0.0	0.0	0.0	0.0	99.9
Recursive with 1/SE(reg) weighting	0.0	0.0	0.0	0.1	0.0	99.9
Recursive with (1+log(t)) weighting	0.0	0.0	0.0	0.1	0.0	99.9
2/3 - 1/3 rule (no weighting)	0.0	1.9	0.8	1.1	1.1	95.2

Table 6: Summary Results for Tests of the Size and Power of the Diebold-Mariano Test

Panel A: Size of DM Test						
	From k+1 (no weight)	t-weight	1/f-weight	1/SE(reg) weight	(1+log(t)) weight	2/3 - 1/3,no weight
T=15	11.64	11.68	11.81	11.23	12.09	25.27
T=25	10.78	10.80	11.05	10.46	11.39	15.71
T=50	9.87	9.89	9.98	9.77	10.37	12.19
T=100	9.68	9.75	9.85	9.68	10.19	11.31
T=250	9.77	9.80	9.87	9.91	10.01	10.52
Panel B: Power of DM Test						
	From k+1 (no weight)	t-weight	1/f-weight	1/SE(reg) weight	(1+log(t)) weight	2/3 - 1/3,no weight
T=15	41.73	41.87	38.94	39.55	41.97	-
T=25	52.46	52.60	47.33	53.92	53.11	-
T=50	62.69	62.95	61.19	64.25	64.44	-
T=100	69.30	69.58	68.87	70.86	71.40	-
T=250	71.91	72.11	71.05	73.62	73.19	-

Notes: Nominal test size: 10%; asymptotic critical values employed; 2/3-1/3 framework not employed in power analysis due to severe over-sizing.

Table 7: Power of Diebold-Mariano Test

DGP	**Comparator Model**	**From k+1 (no weight)**	**t-weight**	**1/f-weight**	**1/SE(reg) weight**	**(1+log(t)) weight**
RWD	AR(4)	62.77	64.19	75.73	70.42	74.25
AR(4)	Linear structural	90.68	91.34	92.95	95.97	95.75
Linear structural	Quadratic Structural	60.18	62.45	67.94	75.90	79.59
Quadratic Structural	Multiplicative Structural	53.15	55.06	59.45	68.00	71.80
Multiplicative Structural	Piecewise linear	6.22	6.20	8.21	9.35	6.59
Piecewise linear	RWD	7.48	7.63	11.24	8.69	8.46
RWD	Linear structural	100.00	100.00	100.00	100.00	100.00
AR(4)	Quadratic Structural	99.97	99.98	99.99	99.93	99.99
Linear structural	Multiplicative Structural	99.98	99.98	99.99	99.90	99.99
Quadratic Structural	Piecewise linear	94.96	95.34	97.07	90.32	96.52
Multiplicative Structural	RWD	24.47	24.43	18.40	29.30	24.25
Piecewise linear	AR(4)	20.22	20.19	14.99	21.69	20.20
RWD	Quadratic Structural	99.99	99.99	100.00	99.89	100.00
AR(4)	Multiplicative Structural	99.37	99.42	99.63	99.12	99.77
Linear structural	Piecewise linear	93.06	93.50	95.16	91.00	96.29
Quadratic Structural	RWD	98.88	98.96	99.34	98.60	99.68
Multiplicative Structural	AR(4)	24.29	24.42	18.32	28.22	24.23
Piecewise linear	Linear structural	18.04	18.15	13.38	19.33	17.82
RWD	Multiplicative Structural	100.00	100.00	100.00	99.99	100.00
AR(4)	Piecewise linear	99.84	99.84	99.93	99.70	99.94
Linear structural	RWD	97.19	97.47	98.43	95.74	98.84
Quadratic Structural	AR(4)	94.56	94.75	97.16	91.62	96.73
Multiplicative Structural	Linear structural	46.07	46.19	36.87	43.56	45.49
Piecewise linear	Quadratic Structural	19.12	19.12	14.06	20.63	18.89
RWD	Piecewise linear	100.00	100.00	100.00	100.00	100.00
AR(4)	RWD	99.99	99.99	99.99	99.99	99.99
Linear structural	AR(4)	99.92	99.92	99.86	99.88	99.91
Quadratic Structural	Linear structural	99.22	99.21	98.83	98.05	99.08
Multiplicative Structural	Quadratic Structural	38.45	38.52	27.28	41.12	37.67
Piecewise linear	Multiplicative Structural	31.05	31.17	21.78	29.87	30.15

Notes: T=100; nominal size of test = 10%.

Table 8: Diebold-Mariano Tests using Meese-Rogoff Exchange Rate Data

		Panel A: GBP-USD					
	Forward rate	Random walk with drift	AR(12)	Hooper-Morton	Frenkel-Bilson	Dornbusch-Frankel	VAR(2)
Recursive from k+1 (no weighting)	1.587	0.184	1.214	7.880	8.200	8.784	3.251
	(0.112)	(0.854)	(0.225)	(0.000)	(0.000)	(0.000)	(0.001)
Recursive with t-weighting	2.135	0.365	1.336	6.609	7.369	7.142	2.892
	(0.033)	(0.715)	(0.181)	(0.000)	(0.000)	(0.000)	(0.004)
Recursive with 1/f-weighting	0.824	0.474	1.267	6.690	7.128	7.580	3.577
	(0.410)	(0.636)	(0.205)	(0.000)	(0.000)	(0.000)	(0.000)
Recursive with 1/SE(reg) weighting	2.293	0.610	2.405	8.969	9.503	9.441	3.849
	(0.022)	(0.542)	(0.016)	(0.000)	(0.000)	(0.000)	(0.000)
Recursive with (1+log(t)) weighting	1.740	0.271	1.222	7.512	7.903	8.306	3.443
	(0.082)	(0.787)	(0.222)	(0.000)	(0.000)	(0.000)	(0.001)
2/3-1/3 (no weighting)	2.037	-1.135	1.138	6.214	6.366	5.947	0.193
	(0.042)	(0.256)	(0.255)	(0.000)	(0.000)	(0.000)	(0.847)
		Panel B: DEM-USD					
	Forward rate	Random walk with drift	AR(12)	Hooper-Morton	Frenkel-Bilson	Dornbusch-Frankel	VAR(2)
Recursive from k+1 (no weighting)	1.003	0.743	1.048	5.202	5.575	5.709	3.047
	(0.316)	(0.457)	(0.295)	(0.000)	(0.000)	(0.000)	(0.002)
Recursive with t-weighting	0.714	1.152	1.040	4.504	4.722	5.041	4.126
	(0.475)	(0.249)	(0.299)	(0.000)	(0.000)	(0.000)	(0.000)
Recursive with 1/f-weighting	0.503	0.291	1.039	6.063	6.874	5.292	3.232
	(0.615)	(0.771)	(0.299)	(0.000)	(0.000)	(0.000)	(0.001)
Recursive with 1/SE(reg) weighting	1.550	0.559	1.390	7.346	8.591	6.782	3.097
	(0.121)	(0.576)	(0.164)	(0.000)	(0.000)	(0.000)	(0.002)
Recursive with (1+log(t)) weighting	0.810	0.993	1.038	4.884	5.226	5.426	3.908
	(0.418)	(0.321)	(0.299)	(0.000)	(0.000)	(0.000)	(0.000)
2/3-1/3 (no weighting)	0.691	1.680	2.199	4.837	4.831	5.802	2.622
	(0.489)	(0.093)	(0.028)	(0.000)	(0.000)	(0.000)	(0.009)
		Panel C: JPY-USD					
	Forward rate	Random walk with drift	AR(12)	Hooper-Morton	Frenkel-Bilson	Dornbusch-Frankel	VAR(2)
Recursive from k+1 (no weighting)	1.196	0.656	1.002	6.574	5.062	6.279	1.300
	(0.232)	(0.512)	(0.316)	(0.000)	(0.000)	(0.000)	(0.194)
Recursive with t-weighting	1.260	0.835	1.005	5.747	4.696	5.492	1.643
	(0.208)	(0.403)	(0.315)	(0.000)	(0.000)	(0.000)	(0.100)
Recursive with 1/f-weighting	0.336	0.063	1.002	5.066	4.039	4.901	-3.887
	(0.737)	(0.950)	(0.316)	(0.000)	(0.000)	(0.000)	(0.000)
Recursive with 1/SE(reg) weighting	1.172	0.218	2.244	6.947	5.031	6.378	1.081
	(0.241)	(0.828)	(0.025)	(0.000)	(0.000)	(0.000)	(0.280)
Recursive with (1+log(t)) weighting	1.169	0.657	1.003	6.289	4.926	6.006	1.370
	(0.242)	(0.511)	(0.316)	(0.000)	(0.000)	(0.000)	(0.171)
2/3-1/3 (no weighting)	1.478	1.309	0.989	6.019	3.893	5.868	2.100
	(0.140)	(0.191)	(0.323)	(0.000)	(0.000)	(0.000)	(0.036)

Note: the models in each column are compared with a random walk; p-values from asymptotic 2-sided critical values are given in parentheses.